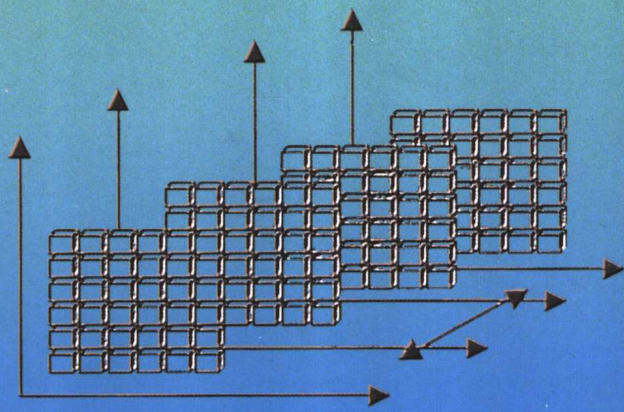
当代情报学(信息管理)前沿丛书

教育部博士点基金资助项目（项目号：01JB870003）

数据挖掘理论与技术

苏新宁　杨建林
邓三鸿　周　军　著



科学技术文献出版社

数据挖掘理论与技术

[redacted]

[redacted]

当代情报学(信息管理)前沿丛书

教育部博士点基金资助项目(项目号:01JB870003)

数据挖掘理论与技术

苏新宁　杨建林　邓三鸿　周　军　著

科学技术文献出版社

Scientific and Technical Documents Publishing House

北　京

图书在版编目(CIP)数据

数据挖掘理论与技术/苏新宁等著.-北京:科学技术文献出版社,2003.6

(当代情报学(信息管理)前沿丛书)

ISBN 7-5023-4273-7

Ⅰ.数… Ⅱ.苏… Ⅲ.情报检索-数据处理 Ⅳ.G354.42

中国版本图书馆 CIP 数据核字(2003)第 020561 号

出　　版　者:科学技术文献出版社

地　　址:北京市复兴路 15 号(中央电视台西侧)/100038

图书编务部电话:(010)68514027,(010)68537104(传真)

图书发行部电话:(010)68514035(传真),(010)68514009

邮购部电话:(010)68515381,(010)68515544-2172

网　　址:http://www.stdph.com

E-mail:stdph@istic.ac.cn;stdph@public.sti.ac.cn

策划编辑:宋振峰　郭伟平

责任编辑:平　平

责任校对:唐　炜

责任出版:王芳妮

发　　行　者:科学技术文献出版社发行　全国各地新华书店经销

印　　刷　者:三河市富华印刷包装有限公司

版(印)次:2003 年 6 月第 1 版第 1 次印刷

开　　本:850×1168　32 开

字　　数:319 千

印　　张:12.75

印　　数:1～3000 册

定　　价:22.00 元

(京)新登字 130 号

内 容 简 介

本书是一部深入探讨数据挖掘理论与技术的著作，全书着重讨论数据挖掘技术在信息处理方面的应用。全书共分七章，详细叙述了数据挖掘的起源及基本概念；数据挖掘的功能过程和方法；基于文本的挖掘技术；内容为 Web 挖掘技术与方法；复杂类型的挖掘技术及数据挖掘在不同领域的应用案例等。可作为广大情报学研究人员和企事业单位专业人员进行数据挖掘研究的参考用书。

科学技术文献出版社是国家科学技术部系统惟一一家中央级综合性科技出版机构，我们所有的努力都是为了使您增长知识和才干。

当代情报学(信息管理)前沿丛书

编　委　会

(按姓氏笔划为序)

丛书总序

梁战平

情报学是一门发展中的新学科，国内外不同时期从不同侧面对它的定义和内涵有各种表述。综合其共同点，情报学是研究有效地运用信息、知识和情报的规律性的一门科学。情报学发源于图书馆学和文献学，已发展成为自然科学、技术科学和社会科学的交叉学科。

1. 信息与情报

(1)“信息链”

信息与情报是情报学的核心问题。什么是信息？什么是知识？什么是情报？对这些基本概念如果没有明确的认识，就不可能获得对情报学及其相关学科的科学理解。“信息”和“情报”，英语都是“Information”。英语的 Information 是一个连续体的概念，“信息链”由 Facts(事实)→Data (数据) →Information (信息) →Knowledge (知识) →Intelligence(“情报”、“智能”)五个链环构成。简单地说，“事实”是人类思想和社会活动的客观映射。“数据”是事实的数字化、编码化、序列化、结构化。“信息”是数据在信息媒介上的映射。“知识”是对信息的加工、吸收、提取、评价的结果。“情报”、“智能”则是运用知识的能力。换句话说，“事实”、“数据”、“信息”、“知识”、“情报”五个链环组成“信息链”(Information

Chain)。在“信息链”中,“信息”的下游是面向物理属性的,上游是面向认知属性的。作为中心链环的“信息”既有物理属性也有认知属性,因此成为“信息链”的代表称谓。

(2)“三个世界”模型

英国科学哲学家卡尔·波普尔(K·Popper)提出的“三个世界”的理论,从哲学高度阐述了信息的属性。波普尔认为,信息有“三个世界”:第一世界是物理领域,第二世界是主观现实领域,第三世界是客观知识领域。根据这个理论,信息分为三大类:第一类是有关客观物理世界的信息,即本体论意义上的信息,它反映事物运动的状态及其变化的方式;第二类是有关人类主观精神世界的信息,即主体论或认识论意义上的隐性信息,它反映人类能感受的事物运动状态及其变化方式,处于意识、思维状态;第三类是有关客观意义上概念世界的信息,即主体论或认识论意义上的显性信息,它反映人类所表述的事物运动状态及其变化方式,用语言、文字、图像、影视、数据等各种载体来表示,汇成一个实在的自主的“信息世界”。以“三个世界”的理论来研究信息、知识、情报,它们之间存在以下关系:

并列关系。事实——数据——信息——知识——情报。

转化关系。数据不会自动变成信息,信息也不会自动变成知识,数据、信息、知识同样也不会自动变成情报。实现从数据到情报的关键要素是人。是人通过信息组织与管理,知识组织与管理来实现信息、知识、情报相互转化。知识本身也是一种信息,情报本身也是一种信息,相互之间可以转化。但是,知识、情报不是一般的信息,而是体现人的认知因素而且在运用中能改变人的行为的特殊信息。

包含关系。信息存在于全部的三个世界中(主观世界、客观的物理世界、客观的概念世界),知识存在于主观世界和客观的概念世界,但不存在于客观物理世界中,因此知识包含于信息之中。情

报也存在于主观世界和客观的概念世界中，是活化了的知识信息，包含于知识、信息之中。

层次关系。从数据提升到信息，主要是对数据之间建立相关性，使其有序化和结构化。从信息提升到知识，主要根据信息的相关性、有序性，进行比较、分析、综合和概括，从中发现问题的本质。从数据、信息、知识提升到情报，主要是采取各种有效的手段和方法激活它们，满足用户的需求。

2. 情报学研究范式

情报学的多学科特性，正是由情报学的多种研究范式决定的。围绕情报学理论研究，可归纳以下研究范式：

(1)机构范式(Institution paradigm)

机构范式是一种视图书馆和情报中心为社会机构的一组思想和观念，以社会学和教育学观点研究图书馆，从图书馆实践出发，研究资料(采集文献)、组织(行政机构和人员管理)、知识属性(分类、编目、采编政策等)，从而驱动资料和组织的有效管理以发挥机构的社会功能。我国20世纪60～70年代情报学以及所探讨的文献合理布局、情报所的地位、作用以及情报政策、管理等都是从机构范式出发，对本行业的问题进行研究。

(2)信息运动范式(Information movement paradigm)

该范式起始于申农和维纳《通信数学理论》一书的通信数学模式：即信息源——传输器——噪音——接受器——信息端。信息运动范式关注的是信息运动的过程——反馈和控制。它构成了当代情报检索系统和文献计量学研究的基础。显然，通信数学模式的概念不适合应用在信息语义上，情报用户被视为情报检索系统以外的被动接受者，要去适应检索系统，利用现有的信息。因此，该范式只是从系统角度去对待情报用户，而不是从情报用户角度了解用户的情报需求。

(3)解释学范式(Hermeneutics paradigm)

伽尔默尔提出解释学的依据是人对信息、情报的解读、解释因人的知识与经验的不同而取舍,因此要研究传播、语言、文字、知识、理解及解释。如果说卡尔·波普尔偏向把情报作为静态的客观知识来加以纯技术性的分析和处理,伽尔默尔的解释学认为,社会文化以及情报消费主体的知识结构和心理状态在查询、解读和利用情报的过程中产生了至关重要的作用,因此必须关注情报流动过程中情报客体与情报消费主体的交融。

(4)技术主导范式(IT-centered paradigm)

V·布什关于实现情报检索自动化的构想,使情报学研究的主流向着利用技术解决问题的范式演变,技术范式对情报学的发展产生了深刻影响。计算机技术突破了人类生产、处理和存贮信息的能力在数量、时间和智力等方面的限制,通信技术的进步,突破了人类传递信息的能力在距离和时间两方面的限制,信息内容开发从点(字、词)、线(字符串、全文文本)、面(数据库、关系数据库)、立体(信息流、物流、资金流的结合)、三维空间(A/V、数据挖掘)到万象空间(虚拟真实)不断纵深发展。情报学研究致力于发展各种先进、高效的情报系统和信息技术应用,但是,情报技术的应用并不是情报学的全部内容,不但如此,由于过分夸大技术的作用,反而导致了重技术轻理论的倾向,忽略情报学的整体研究。

(5)认知范式(Cognitive paradigm)

由于认知科学的发展,一些研究者开始从认知过程,如注意、知觉、表象、记忆、思维、语言等,来观察信息和情报现象。认知范式强调人的知识结构,研究人的信息处理原理,关注情报的利用和吸收,目的是支持和改善情报系统的设计和情报服务。认知观的变迁意味着情报学研究主体从情报检索系统的设计和开发扩大到强调情报用户的知识结构、认知过程、情报行为和人机交互等认知范围。

(6)知识主导范式(Knowledge-based paradigm)

传统情报学的研究对象是文献单元而不是知识内容。英国情报学家布鲁克斯1980年提出了著名的布鲁克斯基本方程式,明确地指出情报学的任务是探索和组织客观知识,情报学要对客观知识进行分析和组织,以便绘制出知识的"认识地图"并最终按"认识地图"来组织知识。情报学从文献层次向知识层次的深化、演进与发展是情报学研究的新趋势。知识有显性知识和隐性知识之分。显性知识存在于信息载体上,通常经过符号化、编码化或结构化等文献处理,内容是固定的,外在的。隐性知识存在于人的大脑中、行为上及概念里,是个人的、没有经过文献化,内部化的,以经验为基础的。隐性知识比显性知识更能激活灵感和启发创新,是一种更有价值的知识,但以往这类知识只能靠个人交流获取,无法收集和加工利用。情报学要超越显性知识,研究收集、筛选、加工、整理隐性知识的理论和规律。当前知识经济、知识组织、知识管理、知识发现、数据挖掘、知识产权保护等问题的研究正在成为情报学界研究热点和学科体系成长的标志,最终将使情报学成为研究知识与知识活动包括知识的激活、扩散、转移、组织、增值、吸收、利用等规律性的一门学科。

(7)经济学范式(Economic paradigm)

情报学与经济学的联系早期仅仅只是引入经济学中的效用、效益等概念、成本—收益分析方法、投入—产出分析方法等基本方法,借用政治经济学的生产—交换—分配—消费模式来评价情报服务的成本与效率。随后,情报的价值、情报传递的成本与效益以及情报工作的效率等也成为情报经济学的主要议题。1979年在荷兰海牙召开了国际情报经济学年会,内容主要围绕情报商品与情报市场研究、情报经济效益研究、情报经济管理研究、情报产业和信息化社会发展研究等方面。面向21世纪,信息经济学的研究方兴未艾,网络革命掀起的全球信息化所提出的众多理论课题与

实践课题正在推动情报经济学开拓新的领域。例如,信息(情报)经纪业、'竞争情报'、博奕论、微观经济学中市场结构理论等,都成为情报经济学研究热点。

(8)人文范式(Culture paradigm)

以人为本的思想必然要同人文科学这一更高层次的概念进行整合,从而研究信息民主与信息专制、信息自由与信息保护、信息平等与信息歧视、信息富裕与信息贫穷、信息共享与信息垄断以及信息污染、信息灾害、信息伦理、信息法律、信息政策、信息文化等以人为主体的信息环境中人与人、人与社会、人与文化的相互关系。突出人文因素的研究,提高人的信息素养,将使情报学更加符合信息化时代特征和情报学自身的发展要求。

3. 国内外情报学发展现状

20 世纪 80~90 年代以来,情报学研究范式的多元化,拓展了情报学研究视野和研究内容,使情报学研究带有时代特征,同信息科学群的其他学科协调、融合、互补,进入了一个情报学整体更新的发展阶段。信息技术是情报学创新的原动力,但国外情报学研究迅速改变"技术至上"的倾向,技术与理论并重,技术与人文并重,技术与经济并重,不断探索情报与技术最佳匹配模式。情报学研究从强调信息需求和信息利用,重视以用户为中心来设计信息系统和情报检索开始,逐步引入解释学、认知观、人文因素等新成分,现在关注的焦点移向知识管理和利用、以人为本、用户/信息/技术/社会和谐共处的生态平衡。情报学不断对传统观念提出质疑,与时代的要求俱进,与技术的发展俱进,与社会的进步俱进,不断拓宽情报学研究领域和研究内容,目前已形成为一门多范式交叉、多学科集成的全方位情报学。

中国情报学研究在 80 年代掀起了两个高潮。一个高潮是引进国外情报理论,开始学习和探讨波普尔的 3 世界理论、布鲁克斯

的知识方程式以及系统论、信息论、控制论、耗散结构论、协同论等，为我国情报学基础理论研究打下基础，一些有影响的情报学专著如《情报学概论》、《情报研究方法论》、《文献计量学》、《情报数学》等相继问世。另一个高潮是开始计算机情报检索的试验、应用和研究，出现了计算机编制主题表、汉字切分、中文全文检索、自动标引等应用研究。中国情报学关注领域和研究重点开始从文献转向技术，从理论转向应用。截止 1998 年统计，新中国成立 50 年来情报学领域发表论文计 18 369 篇，按 11 个论文主题分类，论文数排名分别是情报组织管理、情报基础理论、情报检索、情报分析研究、情报服务、情报搜集、情报技术、情报事业、国外情报事业、情报整理、情报教育。关于理论研究方面，情报学界出版了《现代情报学理论》等专著，近年来在面向二十一世纪的情报学、情报学研究的定量化、情报学认知观、经济情报学、知识组织和管理、竞争情报、内容开发等广泛领域也出现了许多有影响的论文，说明中国情报学研究有新的发展。据 2000 年 9 月统计，中国目前培养情报学硕士的高等院校和情报中心有 22 个；培养情报学博士单位有 4 个；情报学作为一级学科单位有北京大学、武汉大学等 2 个。

4. 情报学与相关学科

(1)情报学与图书馆学、文献学

美国学者 S. Herner 1984 年在“JASIS”上发表的《情报学简史》认为，情报学是在图书馆学、计算机和穿孔卡片、研究与发展、文献学、文献与索引技术、传播学、行为科学、微观与宏观出版、视频与光学等学科领域相互整合的结果。情报学与图书馆、文献学在学科性质上许多共同之处，都要研究编目与分类、存档与索引、检索与获取等技术。图书馆学和文献学是情报学的基础之一。图书馆学是以图书期刊为对象，以馆藏、出纳、阅览等为工作重点，文献学以文献为对象，以揭示报道、加工、研究、提供每篇文献以至每

个数据的内容为重点。情报学以信息和知识为对象,以内容开发利用为重点,广泛采用情报技术产生、搜集、整理、检索、传递、分析、利用情报。情报学对信息加工组织有质的飞跃,对组织信息是由线性组织(字符串、全文文本)、平面组织(数据库、关系数据库)到立体组织(A/V 数据),进而到虚拟组织(虚拟真实,时空信息)。

(2)情报学与信息科学群

信息科学群的崛起,是信息现象日趋复杂化、信息爆炸性增长、知识重要性增加、信息技术飞速发展等因素相互作用的结果。不同学科领域对信息现象的共同探索,形成了信息科学群。信息科学群是以信息为基本研究对象,以信息运动规律和应用方法为主要研究内容,以扩展人类信息功能为中心研究目标而形成的一个横断性、综合性学科群体。情报学是信息科学群的一个分支学科,起着重要作用,为信息科学群各个范畴提供新思路、新概念和新方法。综合有关研究,信息科学群的研究范围包括:哲学范畴、认知范畴、计算机科学范畴、信息交流与管理范畴、社会科学范畴、自然科学和工程技术领域有关信息范畴等。

(3)情报学与信息管理学

情报学与信息管理学具有血缘关系和学科延续性,信息管理学在广度上超过了情报学,而在深度上则逊于情报学。二者之间不是一种取代关系,而是一种衔接关系。从发展趋势看,两者将形成互补互动的学科关系。情报学 50 多年的发展形成的研究方法体系可为信息管理学研究方法体系的建立提供借鉴。信息管理学开发和利用当代信息资源的新技术和方法可为情报学弥补学科空缺领域提供借鉴。对于情报学和信息管理学来说,一方的研究向另一方研究领域发展会给双方学科带来新的研究领域和新的研究方向。

5. 情报学核心研究内容

情报学应该有自己的核心研究内容。情报学作为信息科学群一门独立的学科，必须阐述信息现象并回答有关信息查寻过程中的智力行为问题，而且这种回答必须是科学的并基于在一定程度上是本领域独有的调研方法。ASIS 主席萨拉塞维克(Saracevic)认为，情报学科分为两大块：情报分析和情报检索。情报分析是指：情报学家对文献和文献结构的分析研究，研究作为内容载体的文本；研究不同群体中的信息传播，尤其是科学传播；情报的社会背景；情报利用；情报搜寻和情报行为；关于情报和相关论题的各种理论。现在情报分析与情报检索之间存在鸿沟，情报学的任务就是填平这道鸿沟。他认为，“待这两端成功相连之际，便是情报学这门学科羽翼丰满之时”。综合萨拉塞维克等学者的观点并从现实出发，情报学的核心研究领域可包括理论方法、信息管理和服务、情报分析、信息检索、知识管理、信息技术应用六个组成部分。核心领域涉及的主要研究内容包括：

(1)理论方法

主要探讨和研究情报的性质、现象和过程、各种理论范式、情报学与相邻科学的关系等学科建设方向。当前尤其需要关注信息与社会进步、信息与经济活动、信息与大众传媒、信息与教育、信息与人文、信息构筑、信息生态、信息政策法规、信息伦理、知识产权、行为科学等课题。

(2)研究方法

需要关注文献计量学、信息计量学、网络计量学、科学计量学以及情报的量化分析、引文分析、文献知识发现等课题。

(3)信息管理

包括信息的收集、整理、存储、传播、分析和服务活动；信息资源开发和利用、信息资源的分类、信息资源管理体系、信息资源共

建共享等;信息生产者与用户的关系;信息系统质量评价等;有关信息格式、内容加工和传输的各种标准和规范等。

(4)信息检索

以信息处理和情报内容加工为主的研究。包括:元数据、界面设计、可视化、主题词表、分类表、概念分类、Web 网站构筑、多媒体检索、跨语言检索、检索策略、搜索引擎等。

(5)知识管理

知识单元、知识存储和管理、知识获取、知识提取、知识发现、知识表述和分类、知识挖掘、自然语言理解、语料库、知识工程应用研究、知识管理与统计学、机器学习、自动推理、问题求解、人类常识和专业知识的分析研究、最佳实践(Best practice)和实践团体(Community of Practice, COP)、协同网等。

(6)情报分析研究

从信息挖掘、抽取,对信息进行分析、加工,提供情报咨询服务,以及其相应的信息系统,如竞争情报(CI)、电子数据处理系统(EDPS)、决策支持系统(DSS)、群体决策支持系统(GDSS)、在线分析处理(OLAP)系统、计算机支持协同工作(CSCW)等。

(7)应用和服务

应用范围包括电子商务、电子政务、在线教育、在线学习、在线保健、在线娱乐、在线金融等。服务范围包括网络接入商(IAP)、网络服务商 (ISP)、网络内容商 (ICP)、应用服务商 (ASP)、网络培训商 (ITP)、系统集成商(SI)、网络咨询等。

(8)技术应用

技术对情报学发展的影响。信息内容技术:信息数字化、全文检索、搜索引擎、多媒体内容检索、自动标引、自动翻译、自动摘要、数据挖掘、文本挖掘、信息提取等。计算机与网络技术支持的知识内容加工和知识吸收、转换等。数字图书馆技术。

(9)信息教育与人才培养

包括数字鸿沟、计算机文明、信息技能、专业结构、人才素质、教育制度、在职培训、继续教育、网络教育、网络学习等课题。

6. 情报学研究方法

(1)社会调查法

情报调查法是指人们在社会情报实践活动中对客观情报情况的调查了解与分析研究方法,是搜索、跟踪、获取和开发利用情报资源的一种基本的、有效的方法。这种方法又可分作直接方法与间接方法两大类,前者主要是用现场观察法,后者又分作访问调查与调查表调查。

(2)引文分析法

研究文献的被使用和被引用,也就是研究质量问题。自60年代初以来,由于"科学引文索引"(SCI)的创办,引文分析法已成为一个有相当深度和广度的情报学分支。对引文这一线索进行研究,可以了解某项发明或技术的应用范围、现状、著作水平、学科发展趋势等。

(3)系统科学方法

从系统论、控制论和信息论出发,主要研究科技情报系统的结构、功能和最优设计,以及解决科技情报系统的最佳运行、实现最优服务等问题。

(4)文献计量法

文献计量是情报学与数学、统计学等相互交叉和结合而产生的研究方法。文献计量研究方法包括布拉德福定律、洛特卡定律,齐夫定律、引文规律、文献老化规律、文献增长与冗余等已形成理论体系。文献计量法开始向其他学科输出、扩散、渗透,利用文献计量统计方法,可以描述和解释许多分布机制相似的社会现象,如收入分布、利润分布、人口分布、不合格元件分布、通信间隔分